CLAIMS

What is claimed is:

1. A method for mirroring data between a plurality of sites, comprising:

maintaining, at a first site of the plurality of sites, a record that identifies which transactions that have been executed at the first site have had their redo information replicated to the other sites of the plurality of sites;

determining a priority value associated with a transaction that is to be performed at the first site, wherein the transaction specifies a modification to a data block;

if the priority value is a first value in a set of possible values, then committing the transaction only after the record indicates that redo information associated with the transaction has been replicated to the other sites of the plurality of sites; and

if the priority value is a second value in said set of possible values, then committing the transaction even though the record does not indicate that redo information associated with the transaction has been replicated to the other sites of the plurality of sites.

2. The method of Claim 1, wherein the first value indicates that the transaction should not be lost if the first site becomes inoperable.

3. The method of Claim 1, wherein the second value indicates the transaction can be lost if the first site becomes inoperable.

4. The method of Claim 1, further comprising the step of:

determining whether all other transactions that have committed before the transaction has committed have had their respective redo information replicated to the

other sites of the plurality of sites by comparing a commit record associated with the transaction to the record.

5. The method of Claim 1, wherein the record is a first record, and the method further comprises the step of:

maintaining, at the first site, a second record that identifies which transactions that have executed at the first site have had their redo information logged to persistent storage at the first site.

6. The method of Claim 1, further comprising the step of:

if the priority value is the second value in the set of possible values, then committing the transaction before the record indicates that the redo information generated by the transaction has been replicated to the other sites of the plurality of sites.

7. The method of Claim 5, further comprising the step of:

if the priority value is the second value in the set of possible values, then committing the transaction after the second record indicates that the redo information generated by the transaction has been stored to persistent storage at the first site.

8. The method of Claim 5, further comprising the step of:

determining which transactions that have executed at the first site have had their redo information logged to persistent storage by comparing a commit record associated with the transaction to the second record.

9. The method of Claim 1, further comprising the step of:

if a particular site of the plurality of sites becomes inoperable, then initiating recovery of the particular site after it is determined that all messages transmitted from the particular site to each other site of the plurality of sites have been received at their destination.

10. The method of Claim 1, further comprising the steps of:

at each site of the plurality of sites, determining if a data structure is to be replicated to each other site of the plurality of sites; and

replicating the data structure to each other site of the plurality of sites unless it is determined that the data structure is not to be replicated to each other site of the plurality of sites.

11. A method for storing data, comprising:

at a first site in a plurality of sites, processing a transaction;

generating in volatile memory information that reflects the processed transaction; and

if said information has not been durably stored before either a data block associated with the processed transaction is durably stored or the data block is transferred to another site of the plurality of sites, then durably storing said information before either the data block is durably stored or the data block is transferred to another site of the plurality of sites.

12. The method of Claim 11, wherein the data block is a first data block, wherein the transaction is a first transaction, the information is a first information, and the method further comprises the steps of:

at the first site, processing a second transaction;

generating in volatile memory second information that reflects the processed second transaction; and

if said first information and second information has not been durably stored before either a second data block associated with the processed second transaction is durably stored or the second data block is transferred to another site of the plurality of sites, then durably storing using a batch process said first information and said second information before either the second data block is durably stored or the second data block is transferred to another site of the plurality of sites.

13. The method of Claim 12, further comprising the step of:

determining whether the batch process has completed durably storing the first information and the second information.

14. A method for mirroring data between a plurality of sites, comprising:

maintaining, at a first site of the plurality of sites, a record that identifies which changes made to one or more data blocks stored at the first site have had associated redo information replicated to the other sites of the plurality of sites, wherein the first site implements a write-ahead logging scheme;

determining if the first site replicates, to the other sites of the plurality of sites, write transactions that are executed at the first site in the order in which the write transactions were issued; and

if the first site does not replicate, to the other sites of the plurality of sites, write transactions that are executed at the first site in the order in which the write transactions were issued, then durably storing a data block, in the one or more

data blocks, associated with a transaction only after the record indicates that any write transactions that have updated the data block at the first site have had their respective redo information replicated to the other sites of the plurality of sites.

15. The method of Claim 14, wherein the record is a first record, and further comprising the steps of:

maintaining, at the first site, a second record that identifies which changes made to the one or more data blocks stored at the first site have had associated redo information logged to persistent storage at the first site; and

if the first site does replicate , to the other sites of the plurality of sites, write transactions that are executed at the first site in the order in which the write transactions were issued, then durably storing the data block after the second record indicates that any write transactions that have updated the data block at the first site have had their respective redo information logged to persistent storage at the first site.

16. The method of Claim 14, further comprising the step of:

releasing a lock associated with the data block after the first record indicates that redo information associated with changes made to the data block has been replicated to the other sites of the plurality of sites.

17. The method of Claim 15, wherein the first site replicates write transactions to the other sites of the plurality of sites asynchronously to the completion of the write transaction at the first site.

18. The method of Claim 14, further comprising the step of:

if a particular site of the plurality of sites becomes inoperable, then initiating recovery of the particular site after it is determined that all messages transmitted from the particular site to each other site of the plurality of sites have been received at their destination.

19. The method of Claim 14, further comprising the steps of:

at each site of the plurality of sites, determining if a data structure is to be replicated to each other site of the plurality of sites; and

replicating the data structure to each other site of the plurality of sites unless it is determined that the data structure is not to be replicated to each other site of the plurality of sites.

20. A method for mirroring data between a plurality of sites, wherein the plurality of sites includes a first site, comprising:

at the first site, durably storing a data block prior to durably storing redo information about changes made to the data block; and

at the first site, durably storing the redo information after the changes have been replicated to the other sites in the plurality of sites.

21. The method of Claim 20, wherein the data block is in a plurality of data blocks, wherein changes made to the plurality of data blocks are performed by transactions issued by a single process, and further comprising the step of:

determining if a set of transactions issued by the single process have completed, wherein the set of transactions made the changes to the plurality of data blocks.

22. The method of Claim 20, wherein the data block is in a plurality of data blocks, wherein changes made to the plurality of data blocks are performed by transactions issued by two or more processes, and further comprising the step of:
determining when the changes have been replicated to the other sites in the plurality of sites.

23. The method of Claim 20, further comprising the step of:
if a particular site of the plurality of sites becomes inoperable, then initiating recovery of the particular site after it is determined that all messages transmitted from the particular site to each other site of the plurality of sites have been received at their destination.

24. The method of Claim 20, further comprising the steps of:
at each site of the plurality of sites, determining if a data structure is to be replicated to each other site of the plurality of sites; and
replicating the data structure to each other site of the plurality of sites unless it is determined that the data structure is not to be replicated to each other site of the plurality of sites.

25. The method of Claim 1, wherein the record identifies which transactions that have been executed at the first site have had their redo information replicated to the other sites of the plurality of sites by identifying a portion of a redo log file, and wherein all transactions reflected in the identified portion of the redo log file have been replicated to the other sites of the plurality of sites.

26. The method of Claim 5, wherein the second record identifies which transactions that have executed at the first site have had their redo information logged to persistent storage at the first site by identifying a portion of a redo log file, and wherein all transactions reflected in the identified portion of the redo log file have been logged to persistent storage at the first site.

27. A machine-readable medium carrying one or more sequences of instructions for mirroring data between a plurality of sites, wherein execution of the one or more sequences of instructions by one or more processors causes the one or more processors to perform the steps of:

maintaining, at a first site of the plurality of sites, a record that identifies which transactions that have been executed at the first site have had their redo information replicated to the other sites of the plurality of sites;

determining a priority value associated with a transaction that is to be performed at the first site, wherein the transaction specifies a modification to a data block;

if the priority value is a first value in a set of possible values, then committing the transaction only after the record indicates that redo information associated with the transaction has been replicated to the other sites of the plurality of sites; and

if the priority value is a second value in said set of possible values, then committing the transaction even though the record does not indicate that redo information associated with the transaction has been replicated to the other sites of the plurality of sites.

28. The machine-readable medium of Claim 27, wherein the first value indicates that the transaction should not be lost if the first site becomes inoperable.

29. The machine-readable medium of Claim 27, wherein the second value indicates the transaction can be lost if the first site becomes inoperable.

30. The machine-readable medium of Claim 27, wherein execution of the one or more sequences of instructions by the one or more processors causes the one or more processors to further perform the step of:

determining whether all other transactions that have committed before the transaction has committed have had their respective redo information replicated to the other sites of the plurality of sites by comparing a commit record associated with the transaction to the record.

31. The machine-readable medium of Claim 27, wherein the record is a first record, and wherein execution of the one or more sequences of instructions by the one or more processors causes the one or more processors to further perform the step of:

maintaining, at the first site, a second record that identifies which transactions that have executed at the first site have had their redo information logged to persistent storage at the first site.

32. The machine-readable medium of Claim 27, wherein execution of the one or more sequences of instructions by the one or more processors causes the one or more processors to further perform the step of:

if the priority value is the second value in the set of possible values, then committing the transaction before the record indicates that the redo information generated by the transaction has been replicated to the other sites of the plurality of sites.

33. The machine-readable medium of Claim 31, wherein execution of the one or more sequences of instructions by the one or more processors causes the one or more processors to further perform the step of:

if the priority value is the second value in the set of possible values, then committing the transaction after the second record indicates that the redo information generated by the transaction has been stored to persistent storage at the first site.

34. The machine-readable medium of Claim 31, wherein execution of the one or more sequences of instructions by the one or more processors causes the one or more processors to further perform the step of:

determining which transactions that have executed at the first site have had their redo information logged to persistent storage by comparing a commit record associated with the transaction to the second record.

35. The machine-readable medium of Claim 27, wherein execution of the one or more sequences of instructions by the one or more processors causes the one or more processors to further perform the step of:

if a particular site of the plurality of sites becomes inoperable, then initiating recovery of the particular site after it is determined that all messages transmitted from the particular site to each other site of the plurality of sites have been received at their destination.

36. The machine-readable medium of Claim 27, wherein execution of the one or more sequences of instructions by the one or more processors causes the one or more processors to further perform the steps of:

at each site of the plurality of sites, determining if a data structure is to be replicated to each other site of the plurality of sites; and

replicating the data structure to each other site of the plurality of sites unless it is determined that the data structure is not to be replicated to each other site of the plurality of sites.

37. A machine-readable medium carrying one or more sequences of instructions for storing data, wherein execution of the one or more sequences of instructions by one or more processors causes the one or more processors to perform the steps of:

at a first site in a plurality of sites, processing a transaction;

generating in volatile memory information that reflects the processed transaction; and

if said information has not been durably stored before either a data block associated with the processed transaction is durably stored or the data block is transferred to another site of the plurality of sites, then durably storing said information before either the data block is durably stored or the data block is transferred to another site of the plurality of sites.

38. The machine-readable medium of Claim 37, wherein the data block is a first data block, wherein the transaction is a first transaction, the information is a first information, and wherein execution of the one or more sequences of instructions by the one or more processors causes the one or more processors to further perform the steps of:

at the first site, processing a second transaction;

generating in volatile memory second information that reflects the processed second transaction; and

if said first information and second information has not been durably stored before either a second data block associated with the processed second transaction is durably stored or the second data block is transferred to another site of the plurality of sites, then durably storing using a batch process said first information and said second information before either the second data block is durably stored or the second data block is transferred to another site of the plurality of sites.

39. The machine-readable medium of Claim 38, wherein execution of the one or more sequences of instructions by the one or more processors causes the one or more processors to further perform the step of:

determining whether the batch process has completed durably storing the first information and the second information.

40. A machine-readable medium carrying one or more sequences of instructions for mirroring data between a plurality of sites, wherein execution of the one or more sequences of instructions by one or more processors causes the one or more processors to perform the steps of:

maintaining, at a first site of the plurality of sites, a record that identifies which changes made to one or more data blocks stored at the first site have had associated redo information replicated to the other sites of the plurality of sites, wherein the first site implements a write-ahead logging scheme;

determining if the first site replicates, to the other sites of the plurality of sites, write transactions that are executed at the first site in the order in which the write transactions were issued; and

if the first site does not replicate, to the other sites of the plurality of sites, write transactions that are executed at the first site in the order in which the write transactions were issued, then durably storing a data block, in the one or more data blocks, associated with a transaction only after the record indicates that any write transactions that have updated the data block at the first site have had their respective redo information replicated to the other sites of the plurality of sites.

41. The machine-readable medium of Claim 40, wherein the record is a first record, and wherein execution of the one or more sequences of instructions by the one or more processors causes the one or more processors to further perform the steps of:

maintaining, at the first site, a second record that identifies which changes made to the one or more data blocks stored at the first site have had associated redo information logged to persistent storage at the first site; and

if the first site does replicate , to the other sites of the plurality of sites, write transactions that are executed at the first site in the order in which the write transactions were issued, then durably storing the data block after the second record indicates that any write transactions that have updated the data block at the first site have had their respective redo information logged to persistent storage at the first site.

42. The machine-readable medium of Claim 40, wherein execution of the one or more sequences of instructions by the one or more processors causes the one or more processors to further perform the step of:

releasing a lock associated with the data block after the first record indicates that redo information associated with changes made to the data block has been replicated to the other sites of the plurality of sites.

43. The machine-readable medium of Claim 41, wherein the first site replicates write transactions to the other sites of the plurality of sites asynchronously to the completion of the write transaction at the first site.

44. The machine-readable medium of Claim 40, wherein execution of the one or more sequences of instructions by the one or more processors causes the one or more processors to further perform the step of:

if a particular site of the plurality of sites becomes inoperable, then initiating recovery of the particular site after it is determined that all messages transmitted from the particular site to each other site of the plurality of sites have been received at their destination.

45. The machine-readable medium of Claim 40, wherein execution of the one or more sequences of instructions by the one or more processors causes the one or more processors to further perform the steps of:

at each site of the plurality of sites, determining if a data structure is to be replicated to each other site of the plurality of sites; and

replicating the data structure to each other site of the plurality of sites unless it is determined that the data structure is not to be replicated to each other site of the plurality of sites.

46. A machine-readable medium carrying one or more sequences of instructions for mirroring data between a plurality of sites, wherein the plurality of sites includes a first site, wherein execution of the one or more sequences of instructions by one or more processors causes the one or more processors to perform the steps of:
at the first site, durably storing a data block prior to durably storing redo information about changes made to the data block; and
at the first site, durably storing the redo information after the changes have been replicated to the other sites in the plurality of sites.

47. The machine-readable medium of Claim 46, wherein the data block is in a plurality of data blocks, wherein changes made to the plurality of data blocks are performed by transactions issued by a single process, and wherein execution of the one or more sequences of instructions by the one or more processors causes the one or more processors to further perform the step of:
determining if a set of transactions issued by the single process have completed, wherein the set of transactions made the changes to the plurality of data blocks.

48. The machine-readable medium of Claim 46, wherein the data block is in a plurality of data blocks, wherein changes made to the plurality of data blocks are performed by transactions issued by two or more processes, and wherein execution of the one or

more sequences of instructions by the one or more processors causes the one or more processors to further perform the step of:

determining when the changes have been replicated to the other sites in the plurality of sites.

49. The machine-readable medium of Claim 46, wherein execution of the one or more sequences of instructions by the one or more processors causes the one or more processors to further perform the step of:

if a particular site of the plurality of sites becomes inoperable, then initiating recovery of the particular site after it is determined that all messages transmitted from the particular site to each other site of the plurality of sites have been received at their destination.

50. The machine-readable medium of Claim 46, wherein execution of the one or more sequences of instructions by the one or more processors causes the one or more processors to further perform the steps of:

at each site of the plurality of sites, determining if a data structure is to be replicated to each other site of the plurality of sites; and

replicating the data structure to each other site of the plurality of sites unless it is determined that the data structure is not to be replicated to each other site of the plurality of sites.

51. The machine-readable medium of Claim 27, wherein the record identifies which transactions that have been executed at the first site have had their redo information replicated to the other sites of the plurality of sites by identifying a portion of a redo

log file, and wherein all transactions reflected in the identified portion of the redo log file have been replicated to the other sites of the plurality of sites.

52. The machine-readable medium of Claim 31, wherein the second record identifies which transactions that have executed at the first site have had their redo information logged to persistent storage at the first site by identifying a portion of a redo log file, and wherein all transactions reflected in the identified portion of the redo log file have been logged to persistent storage at the first site.

53. The method of Claim 14, wherein the record identifies which changes are made to the one or more data blocks stored at the first site have had associated redo information replicated to the other sites of the plurality of sites by identifying a portion of a redo log file, and wherein all changes in the identified portion of the redo log file have been replicated to the other sites of the plurality of sites.

54. The method of Claim 15, wherein the second record identifies which changes are made to the one or more data blocks stored at the first site have had associated redo information logged to persistent storage by identifying a portion of a redo log file, and wherein all changes in the identified portion of the redo log file have been logged to persistent storage.

55. The machine-readable medium of Claim 40, wherein the record identifies which changes are made to the one or more data blocks stored at the first site have had associated redo information replicated to the other sites of the plurality of sites by identifying a portion of a redo log file, and wherein all changes in the identified portion of the redo log file have been replicated to the other sites of the plurality of sites.

56. The machine-readable medium of Claim 41, wherein the second record identifies which changes are made to the one or more data blocks stored at the first site have had associated redo information logged to persistent storage by identifying a portion of a redo log file, and wherein all changes in the identified portion of the redo log file have been logged to persistent storage.